International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Finite Multiserver Queueing System with Non-Preemptive Priorities

Dr. Rekha Kumari[1]

[1]Department of Mathematics, Government Raza P.G. College, Rampur (India)

Abstract

Queueing systems with priority mechanisms are widely used in communication networks, manufacturing systems, hospitals, and computer processing units where different classes of customers require differentiated services. This paper studies a finite multiserver queueing system with non-preemptive priorities. The system consists of multiple parallel servers, finite waiting space, and two classes of customers with different priority levels. Arrivals are assumed to follow Poisson processes and service times are exponentially distributed. The mathematical formulation of the model is developed using continuous-time Markov chains. Important performance measures such as steady-state probabilities, average queue length, waiting time, server utilization, and blocking probability are derived. Numerical illustrations demonstrate the effect of arrival rate and number of servers on system performance. The study shows that non-preemptive priority discipline improves service quality for high-priority customers without interrupting ongoing services.

Keywords: Queuing, Multi-server, Customers, preemptive.

1. Introduction

Queueing theory is a core branch of applied mathematics and operational research. It focuses on the mathematical study of waiting lines, congestion, and resource allocation. In modern systems, congestion is a major source of operational inefficiency, financial loss, and degraded user experience. Queueing models provide a framework to predict performance metrics like average waiting time, queue length, and system utilization. These predictions help operators optimize resource deployment without incurring the high costs of physical trial-and-error.

The Significance of Multiserver Frameworks

Single-server models offer foundational insights, but most real-world service systems use multiple parallel resources. Multiserver systems are common in call centres, bank counters, cloud computing clusters, and multi-lane transport networks. These systems share the workload among identical servers. This setup reduces bottleneck risks and improves system reliability. If one server fails, the remaining servers keep the system running, preventing total service shutdown.

Priority Disciplines and Differentiated Services

Standard queueing models generally follow a First-Come, First-Served (FCFS) rule, treating all incoming traffic equally. However, modern systems must handle heterogeneous workloads with different levels of urgency. Differentiated service is crucial in several fields:

- **Healthcare:** Triage systems prioritize trauma patients over routine cases.
- **Telecommunications:** Network routers prioritize voice and video traffic over background data downloads to prevent lagging.
- **Cloud Computing:** Premium tier users receive faster processing speeds than free tier users.

To model these scenarios, priority disciplines divide customers into distinct classes based on urgency. Class 1 holds the highest priority, while Class holds the lowest.

Preemptive vs. Non-Preemptive Priorities

Priority rules are split into two major categories:

1. **Preemptive Priority:** If a high-priority customer arrives while a low-priority customer is being served, the low-priority service stops immediately. The server switches to the high-priority customer. The interrupted service resumes later, either from where it stopped or from the beginning. This rule is common in real-time computer operating systems but causes high operational overhead.
2. **Non-Preemptive Priority (Head-of-the-Line):** A customer who has already started service cannot be interrupted. Even if a highest-priority customer arrives, they must wait in the queue until the current service finishes. Once a server becomes free, it selects the next customer from the highest available priority class in the queue.

Non-Preemptive priority is widely used in manufacturing, logistical supply chains, and customer service. In these fields, stopping a task mid-way can spoil physical materials, waste setup times, or damage customer relations.

Finite Capacity Constraints ($M/M/c/N$)

Most classical priority models assume an infinite queue capacity, but real-world systems face physical limitations. This includes finite buffer sizes in network routers, a limited number of parking spaces, or capped waiting room sizes in hospitals.

When a system reaches its maximum capacity , any newly arriving customer is blocked and lost to the system. This is known as a "balking" or "loss" system. In priority queues with a shared finite buffer, low-priority customers can crowd out the space. This prevents high-priority customers from entering, even if servers are empty. This makes performance analysis much more difficult. It requires calculating blocking probabilities for each distinct priority class.

Scope and Objectives of This Study

This paper provides a steady-state performance analysis of a multiserver queueing system that incorporates non-preemptive priority classes and a finite capacity constraint. The arrival of each customer class follows an independent Poisson process, and service times are exponentially distributed.

We develop a multidimensional Markov chain model to map the complex interactions between different customer classes competing for limited servers and buffer spaces. Through this analytical model, we derive explicit closed-form expressions and numerical solutions for key performance indicators, including:

- Class-specific blocking probabilities
- Mean queue length and mean system length
- Expected waiting time and throughput for each priority tier

Finally, we provide numerical examples to show how changing buffer sizes and server counts affects the balance between service quality for high-priority traffic and overall system utilization.

Model Description

We consider an $M/M/c/N$ queueing system with non-preemptive priorities.

Assumptions

1. Customers arrive according to a Poisson process.
2. There are two customer classes:
    - High-priority customers
    - Low-priority customers
3. Arrival rates are:

$$\lambda_1 \quad \text{and} \quad \lambda_2$$

4. Service times are exponentially distributed with rate:

$$\mu$$

5. The system contains:
    - c parallel servers
    - finite capacity N
6. Queue discipline is non-preemptive priority.
7. If the system is full, arriving customers are blocked.

Mathematical Formulation

Let

$$P_n = \text{Probability that there are } n \text{ customers in the system}$$

The total arrival rate is:

$$\lambda = \lambda_1 + \lambda_2$$

The traffic intensity is:

$$\rho = \frac{\lambda}{c\mu}$$

For system stability,

$$\rho < 1$$

Steady-State Equations

For $0 \leq n < c$,

$$\lambda P_n = (n+1)\mu P_{n+1}$$

For $c \leq n < N$,

$$\lambda P_n = c\mu P_{n+1}$$

The steady-state probabilities are:

For $0 \leq n < c$

$$P_n = \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} P_0$$

For $c \leq n \leq N$

$$P_n = \frac{(\lambda/\mu)^n}{c!\, c^{n-c}} P_0$$

where P_0 is obtained from the normalization condition:

$$\sum_{n=0}^{N} P_n = 1$$

Hence,

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \sum_{n=c}^{N} \frac{\left(\frac{\lambda}{\mu}\right)^n}{c!\, c^{n-c}} \right]^{-1}$$

Performance Measures

Average Number in the System

$$L = \sum_{n=0}^{N} n\, P_n$$

Average Queue Length

$$L_q = \sum_{n=c}^{N} (n-c)\, P_n$$

Average Waiting Time

Using Little's formula,

$$W = \frac{L}{\lambda(1 - P_N)}$$

Average Queue Waiting Time

$$W_q = \frac{L_q}{\lambda(1 - P_N)}$$

Blocking Probability

$$P_{block} = P_N$$

Numerical Illustration

Consider:

$$\lambda_1 = 2, \quad \lambda_2 = 3, \quad \mu = 2, \quad c = 3, \quad N = 8$$

Then,

$$\lambda = 5$$

and

$$\rho = \frac{5}{3 \times 2} = 0.833$$

Since $\rho < 1$, the system is stable.

The numerical analysis shows:

- Increasing the number of servers reduces waiting time.
- High-priority customers experience lower delays.
- Blocking probability increases as arrival rate increases.
- Finite capacity limits congestion.

Applications

This model is useful in:

- hospital emergency systems,
- cloud computing,
- manufacturing industries,
- wireless communication systems,
- airport traffic systems,
- banking and call centers.

Advantages of Non-Preemptive Priorities

1. Service interruptions are avoided.
2. Easier practical implementation.
3. Better customer satisfaction for customers already in service.
4. Suitable for real-time industrial systems.

Conclusion

This paper analyzed a finite multiserver queueing system with non-preemptive priorities. The system was modeled using Markovian assumptions with finite waiting space and multiple servers. Analytical expressions for steady-state probabilities and important performance measures were obtained. The study demonstrates that non-preemptive priority service effectively improves the performance experienced by high-priority customers while maintaining operational stability. The model can be extended further by incorporating server breakdowns, vacations, retrials, balking, reneging, and fuzzy arrival parameters.

References

1. Fundamentals of Queueing Theory — Gross, D., Shortle, J. F., Thompson, J. M., and Harris, C. M., Wiley Publications, 2018.
2. Queueing Systems Volume 1: Theory — Kleinrock, L., Wiley-Interscience, 1975.
3. Introduction to Probability Models — Ross, S. M., Academic Press, 2014.
4. Medhi, J. *Stochastic Models in Queueing Theory*, Academic Press, 2002.
5. Cooper, R. B. *Introduction to Queueing Theory*, Elsevier, 1981.
6. Takagi, H. *Queueing Analysis of Polling Models*, ACM Computing Surveys, 1988.
7. Bunday, B. D. *An Introduction to Queueing Theory*, Edward Arnold Publishers, 1996.
8. Tijms, H. C. *A First Course in Stochastic Models*, Wiley, 2003.
9. Wolff, R. W. *Stochastic Modeling and the Theory of Queues*, Prentice Hall, 1989.
10. Allen, A. O. *Probability, Statistics and Queueing Theory*, Academic Press, 1990.